DEVELOPING VALID LEVEL 2 EVALUATIONS*

by: Ken Phillips

PHILLIPS ASSOCIATES

[*] Published in *Training Today*, Fall 2009 (A quarterly magazine published by the Chicagoland Chapter of ASTD)

DEVELOPING VALID LEVEL 2 EVALUATIONS

By Kenneth R. Phillips

President, Phillips Associates

Collecting anecdotal information regarding training effectiveness is a start. To establish real credibility and prove value, Workplace Learning and Performance (WLP) professionals, need to measure whether or not participants actually learned something – a Level 2 evaluation. Unfortunately, conducting Level 2 evaluations is not something many WLP professionals do well. Many who are uninformed in the art and science of test creation often develop questions that either contain obvious clues to the correct answer or are overly difficult and discourage participants from getting the right answer. In either case, the result is an invalid Level 2 evaluation – one that doesn't measure what it is supposed to and is either unfair to the learner or to the organization.

> *"Many [WLP professionals] are uninformed in the art and science of test creation [and] often develop questions that either contain obvious clues to the correct answer or are overly difficult and discourage participants from getting the right answer."*

In addition, invalid Level 2 evaluations can put WLP professionals at risk by creating situations where it appears:

1. That learning took place when it didn't (the evaluation contained obvious clues to the correct answers).

OR

2. That learning didn't take place when it did (the evaluation was overly difficult and discouraged learners from getting the right answers).

In the first situation, executives may question why participant job behavior didn't change (Level 3) or business results didn't improve (Level 4) if learning improved. In the second situation, executives may question why time and money was wasted on training if participants didn't learn anything. In either case, your reputation and credibility are on the line and certain to suffer in the eyes of company executives. However, both of these situations can be avoided simply by following a set of proven test creation guidelines and tips that result in the development of valid Level 2 evaluations.

Top Ten Test Creation Guidelines

There are ten guidelines to follow when developing Level 2 evaluations that are fair to both the learner and the organization. These guidelines apply regardless of the types of questions contained in your evaluation -- multiple choice, true/false, matching or fill-in-the-blank.

1. Focus on creating Level 2 evaluations that test for understanding not just knowledge.

(This guideline is courtesy of Matt Allen, an I/O consultant with HumRRO a Washington D. C. based human and organizational performance research-consulting company.) For example, take the following questions:

Question	Comments
• What do the letters TV stand for? • What is the main function of a TV?	Recall focused questions such as these tend to test knowledge, but not understanding. Moreover, if learners merely “know” something, but don’t understand it, they’re also likely to forget it shortly after the learning program is over.
• What physical principal is used to display images on a TV?	While correctly answering this question requires more in-depth knowledge of TVs, it’s still a fact based question.
• Your TV is not working. What’s the most likely cause of the problem given the following symptoms…?	Testing for true understanding requires the use of application type questions such this one.

As you can see, answering each of these questions correctly requires a deeper level of understanding of the topic. Correctly answering application type questions assess whether or not participants really understand the content and, after all, isn’t that what a Level 2 evaluation is really trying to measure?

> *“Correctly answering application type questions assess whether or not participants really understand the content and, after all, isn’t that what a Level 2 evaluation is really trying to measure?”*

2. Where appropriate, use Level 2 evaluations for reinforcement as well as evaluation.

Administering the evaluation at a point either weeks or months after the conclusion of a learning program positions it to serve both as reinforcement and as an evaluation. It also increases the credibility of the results. For example, executives expect participant knowledge scores to improve immediately after a learning program is completed. However, if scores show improvement weeks or months after a program is completed, this shows that participants retained what they learned and adds credibility to the results.

If you’re having difficulty getting executive buy in for your evaluation efforts, positioning them as reinforcement can make it easier to get their support.

Executives often perceive evaluation as something that benefits you, and reinforcement as something that benefits them. Executives expect participant learning to occur or they wouldn’t send their employees to your training in the first place. Therefore, demonstrating that learning took place is seen as adding little value.

Reinforcement activities are perceived as beneficial because they extend the reach of the learning program and help to ensure that the employees who attend your training actually learned something.

3. Group questions by topic or concept for scoring, but randomize for administration.

The reason for this is because sometimes one question on a topic provides a clue to the correct answer to another question on that topic. This can be avoided by randomizing items related to the same topic or concept. However, when scoring the evaluation, you'll want to re-group the questions by topic in order to identify trends such as all or most of the questions on a topic being answered incorrectly. This suggests that either the topic was poorly presented or that particular learners need retraining.

4. All evaluation items should discriminate between participants who know the material taught really well from those who don't.

- Check all multiple choice questions to be sure none of the response choices are being over or under selected. Response alternatives that are over or under selected should be revised either to make the choice less obvious or more attractive.

"Response alternatives that are over or under selected should be revised either to make the choice less obvious or more attractive."

- True/false questions should be rewritten if everyone (or nearly everyone) chooses either True or False and it's the wrong answer. Remember the goal is to create an evaluation that is fair to both the learner and the organization.

5. Avoid compound questions that ask for more than one thing.

(This guideline also is courtesy of Matt Allen.) Learners find them confusing and view them as unfair. In addition, avoid questions with compound answers. For example:

Question	Comments
What do the letters in the acronym ADDIE stand for? • Analyze, Design, Develop, Implement and Evaluate • Analyze, Design, Develop, Integrate and Evaluate • Analyze, Design, Develop, Implement and Execute • Analyze, Design, Deploy, Implement and Evaluate	Learners who are test savvy will pick the first option, which is the correct answer, because Integrate, Execute and Deploy only appear in one option and Analyze, Design, Develop, Implement and Evaluate appear in multiple options.

The solution: make both the question and the options short and to the point.

6. Don't test participants on concepts or material that wasn't covered in the learning program.

(This guideline, along with the next four, are adapted from Nanette Miner in an article published in T + D magazine titled "The Art of Test Creation.")

This may seem obvious, but how many times have you taught a learning program and not covered all the material or only covered some of it in a cursory fashion because you ran short of time? The point is that it's not fair to test participants on material that wasn't covered or was only covered in a cursory fashion during the learning program.

7. Write all test items the same way the material was taught.

Don't ask "null" questions such as: "Which of the following is not one of the steps in the ADDIE model?" Null questions are viewed as tricky and unfair. Why reinforce something you don't want participants to remember?

8. Provide clear test instructions.

If you're administering the evaluation live, have participants read through the instructions first to be sure they understand what is expected. If you're not administering the evaluation live, ask one or two colleagues to read through the instructions to be sure they are clear. Unclear test instructions cause participants to view the evaluation as unfair.

"Don't ask participants to recall information from memory on the evaluation if they don't have to recall the information from memory while performing their job."

9. Allow participants to use test aids during the evaluation, if they use them when performing their job.

Don't ask participants to recall information from memory on the evaluation if they don't have to recall the information from memory while performing their job.

10. Avoid developing evaluation items that contain trivial information.

Trivial information is anything included in the item that isn't needed to understand the question. For example, take the following question:

Chris is an internal workplace learning and performance consultant. She has been asked by her boss, Larry, the VP of HR, to design and deliver a new hire orientation program at four company locations across the U.S., Boston, Dayton, Omaha, and Oakland. Larry has also requested that the program not be longer than four hours. What approach should Chris use to design the training?

The names Chris and Larry, Larry's title and the identification of the four company locations all constitute trivial information. A more concise version of the question might read something like:

You are an internal workplace learning and performance consultant and have been asked by your boss to design and deliver a four hour new hire orientation training program. The sessions will be held at four company locations across the U. S. What approach would you use to design the training?

Test Item Creation Tips

In addition to the test creation guidelines, creating valid Level 2 evaluations that are fair to both the learner and the organization also means paying attention to the following tips when creating multiple choice, true/false, matching and fill-in-the-blank questions. (Many of these tips are also adapted from Nanette Miner's article.)

> *"[Multiple choice questions] are easy to grade and when developed correctly – they contain neither obvious clues to the correct answer nor are overly difficult – are the most valid."*

Multiple choice questions are the most popular type question used when creating a knowledge test. They are easy to grade and when developed correctly – they contain neither obvious clues to the correct answer nor are overly difficult – are the most valid. However, they are difficult to write because there can be only one correct answer. Following is a list of common errors made when creating multiple choice test questions and a tip on how to avoid each one.

Common Errors	Tips
A tendency for the correct answer to be the longest and to sound like a definition.	Make sure all the response alternatives contain about the same number of words and sound similar.
The wording of the question reveals the correct answer. For example: The ADDIE model is used primarily as an: A. Instructional design tool B. Measurement and evaluation tool C. Change management tool D. Process improvement tool	Savvy test takers know that the correct answer to this question has to be "A" because it's the only response alternative that begins with a vowel and is grammatically correct with "an" at the end of the question. If the correct answer begins with a vowel, end the question with a(n). Placing the "n" in parentheses enables any of the response choices to be correct.

Common Errors (cont.)	Tips (cont.)
Some response alternatives are obviously wrong or not plausible thus compromising the validity of the question. For example, take the following: __________________ is a useful technique to consolidate what has been previously discussed and to move the focus of a conversation from topic to topic. A. Arguing B. Interrupting C. Summarizing D. Initiating Note: See overall guideline four above for information on how to check all your response alternatives to be sure they are viewed as plausible.	Savvy test takers know that response choices A and B are obviously wrong. Therefore, the probability of guessing the correct answer, even without possessing the requisite knowledge, increases from 25 percent to 50 percent, thus reducing the validity of the question. Develop only plausible response alternatives. Some techniques for developing plausible alternatives also courtesy of Matt Allen include: 1) Use common misunderstandings or confusions about the program content 2) Use other familiar, but incorrect, phrases or concepts 3) Use common errors made with the program content 4) Skip a step in a multi-step process.
The use of "All the above" or "None of the above" as a response choice. Very often when "All the above" and "None of the above" are used as response alternatives, they also are the correct answer.	When savvy test takers see these response alternatives in some test questions, but not all, they know there is a high probability that it is the correct answer. If you need to use "All the above" or "None of the above" as a response alternative, be sure to include one or the other as an alternative in all or nearly all your multiple choice questions.

"When savvy test takers see these response alternatives in some test questions, but not all, they know there is a high probability that it is the correct answer."

"Savvy test takers know that while it's possible a statement might be always or never true; usually that is not the case."

True/False are the second most common type question used in Level 2 evaluations. They are easy to write, but also tend to be the least valid -- learners have a 50/50 chance of guessing the correct answer. Following is a list of common errors and tips how to overcome each one.

Common Errors	Tips
The tendency to develop more True than False questions. True questions are easier to write thus some WLP professionals tend to develop more True than False questions when creating Level 2 evaluations.	Savvy test takers know that when they're not sure of the correct answer, choosing True often gives them a better chance of guessing the correct answer When creating True/False questions, keep a balance between the number of each.
The development of questions that are not completely True or False. For example: February has 28 days. A. True B. False It's true that February has 28 days, but every four years (leap year) it has 29. Thus, learners could argue, regardless of whether they chose True or False, that their answer is correct. Since only one answer can be correct, learners who chose the other answer are going to see the question as unfair or tricky.	Only create questions that are entirely True or False.
The inclusion of words like "never" and "always" in the item. For example: Open-ended questions are always preferred to close-ended questions. A. True B. False False is the correct answer above.	Savvy test takers know that while it's possible a statement might be always or never true; usually that is not the case. Therefore, when in doubt about the correct answer, savvy test takers will choose False because they know it gives them the best chance of guessing the correct answer. Avoid the inclusion of "absolute determiners" like always and never in your items.

Matching questions, a third type of Level 2 evaluation question, are easy to create because only one "B" column correct answer is required for each "A" column question. In contrast, multiple choice questions require at least three or four plausible answers. However, as with both multiple choice and true/false questions, there are a few common errors made. A description of these and tips on how to prevent them follow:

Common Errors	Tips
Creating a matching question that contains more than 10 "A" column questions or "B" column answers. According to Cognitive Learning Science research, we humans possess the mental capacity to work with seven (plus or minus two) different pieces of information at the same time. Therefore, matching questions with more than 10 "A" column questions or "B" column answers are likely to be perceived by learners as overwhelming and unfair.	Keep the number of items in both the "A" and "B" columns to 10 or fewer. If there are more than 10 items on a particular topic that you want to include in a matching question, then break up the items into chunks of 10 or fewer.
Including more than three extra "A" column questions than "B" column answers or vice versa. While having more questions than answers or more answers than questions increases the difficulty level of a matching question, it's important to limit the number of extra questions or answers so as not to overwhelm the learner. Having too many extra questions or answers is likely to be perceived by learners as tricky and unfair.	Limit the number of extra questions or answers to no more than three. Limiting the number to three or fewer ensures that your matching questions will be fair to both the learner and the organization.

"While having more questions than answers or more answers than questions increases the difficulty level of a matching question, it's important to limit the number of extra questions or answers so as not to overwhelm the learner."

Fill-in-the-Blank questions, the fourth type of Level 2 evaluation question, are used to test for learner recall of key facts and concepts. They are easy to create, but more time consuming to grade than multiple choice, true/false or matching questions. They also don't test for understanding. Note: See overall guideline one above for the difference between testing for recall and testing for understanding.

Common Errors	Tips
Creating fill-in-the-blank questions that ask learners to recall obscure facts and concepts. When this occurs, learners view the question as tricky and unfair.	When creating fill-in-the-blank questions, be sure the facts and concepts you're asking learners to recall are important to know.

In summary, Level 2 evaluations often miss the mark because they may not be developed by someone informed in the art and science of test creation. Often questions either contain obvious clues to the correct answer or are overly difficult. In either case, the result is an invalid Level 2 evaluation. However, by following the guidelines and tips described above, you'll be able to create valid Level 2 evaluations that measure what they purport to and are fair to both the learner and the organization, which after all is the true purpose of a Level 2 evaluation anyway.

"...by following the guidelines and tips described above, you'll be able to create valid Level 2 evaluations that measure what they purport to and are fair to both the learner and the organization..."

For more information, contact:
Ken Phillips
Phillips Associates
34137 N. Wooded Glen Drive
Grayslake, IL 60030
(847) 231-6068
www.phillipsassociates.com
ken@phillipsassociates.com